Trends in scientific document search

Stefan Geißler
Expert System Deutschland GmbH

Trends in scientific document search

One of the ironic observations regarding search today is that in many large corporate environments, search capabilities used by professionals in their daily work are lightyears behind what teenagers today take for granted when using popular platforms on the internet (Google, Amazon, Spotify and the like): Typo-tolerant search, semantic abstraction across synonyms or subterm-superterm relations are still largely absent in many places. As a lot of corporate scientific search requires confidentiality (the company may not want the topics of their searches to become public, let alone the company-internal documents on which it is performed) public platforms dedicated to scientific search such as Google Scholar, Semantic Scholar (www.semanticscholar.org/) or PubMed are not always the complete answer.

This short essay discusses some ingredients to (scientific) document search that should be assessed and considered when planning to update a search environment.

Semantic abstraction, allowing to bridge the gap between the terms used in the user's query and the terms in the relevant documents is perhaps the most beneficial extension beyond simple string matching that search should offer: A user looking for information on "laptops" expects to find matches also when they talk about "notebooks"; what is called "lesion" in some documents might be called "injury" in others. Accounting for these term relations by means of adding a thesaurus is a well-established practice in many domains: A search for a given term is extended automatically to this term's synonyms or subterms. Especially in the medical field, resources like the Medical Subject Headings Thesaurus (NLM's MeSH, www.nlm.nih.gov/mesh) are used extensively to facilitate search. Defining and maintaining a large thesaurus however is a complex project and many smaller domains lack a thesaurus like MeSH to this date.

A relatively new approach to allow search environments to handle term mismatches are the so-called word embeddings [1]. Word embeddings allow to assess term term similarities by comparing the typical contexts of terms as observed in large document collections. Since this does not require manual annotation, just raw text, and since efficient and free implementations of the respective algorithms exist (e.g [2]), word embeddings have become very popular and have established themselves as a kind of de facto standard processing steps for many NLP-related tasks. Pre-computed resources (i.e. the vectors that represent the word embeddings) trained on huge public corpora are freely available (e.g. [3]) and they can with moderate computing power be extended with (or calculated from scratch on) one's own document collections. For a more in-depth introduction to the concept of word x embeddings see [4].

While word embeddings exhibit some striking properties [5], the NLP community sometimes jokingly declares that it is almost illegal to talk about word embeddings without mentioning the famous "king – man + woman = queen" example (that shows that using vector representations, lexical semantics can to some extent be expressed as vector algebra) it is also important to be aware of the limitations of the approach: Word embeddings are good at detecting word similarities but they often have a hard time distinguishing different kinds of relatedness: a term like "diabetes" has a paradigmatic relation to terms like "obesity" or "Crohn's disease" in that all these are medical conditions whereas it has syntagmatic relations to terms like "insulin" or "sugar" in that these terms tend to cooccur with "diabetes". A user searching for "diabetes" however, might be confused to see her query extended in the background with "Crohn's disease". Word embeddings in search environments must be used with care in order to account for these effects [6].

Word embeddings are an impressive approach to semantic word relations but the requirement to enhance the reach and accuracy of scientific search doesn't stop on the word/term level: Many relevant questions that keep users busy, are concerned with specific relations between concepts and entities. An information demand such as "Show me evidence where the administration of estradiol to women of age 50 and beyond lead to decreased bone mineral density!" involves a host of analysis requirements that are way beyond term matching approaches: Properly handling this query would need to account for find a relation between the "administration" and the administered substance "estradiol" as well as between the administration and the observed effect (reduced bone mineral density). Search requirements like that can be interpreted as textual entailment tasks (find documents where the content entails the relations expressed in the query. It doesn't come as a surprise that also on this type of tasks, deep-learning inspired approaches have led to impressive progress recently: The best reported results on the SNLI corpus [7] with ~90% accuracy have been obtained by a sophisticated neural network [8].

These results are highly impressive, given that they address a complex task (deciding whether or not a sentence is semantically entailed in another or not) without prior and manually coded world knowledge. Yet the results are possible largely thanks to the huge SNLI corpus of more than half a million of hand annotated training samples.

Preparing training corpora in commercial projects on new tasks, however, often requires considerable resources in time and money and therefore makes the application many machine learning approaches challenging. There is reason to assume that task-specific approaches to complex search requirements will continue to benefit from NLP-inspired methods. An example of this NLP-driven search environment is the work done at Semiring [9] where legal documents are analyzed, collecting relations between the involved concepts and entities and the resulting collection of facts is fed into a graph database for later search and inference. Ontological knowledge (the CEO of a company has to be of type human) can be added and used to flag conflicting assertions and resolve ambiguities.

Regarding the title of this issue of this publication "Surfing and drilling in the modern scientific world" we can conclude that often both is necessary: Surfing where a user is taken from an initial concept to related topics he or she may not initially have had in mind, as well as drilling, where with the help of both quantitative as well as symbolic methods, searches can be made more complete and more focused at the same time. One exciting aspect of today's landscape around these topics is the immense wealth of established methods, algorithms, libraries and resources that are available to jump start specific search projects: State of the art deep learning libraries (Keras, Torch, TensorFlow), powerful NLP platforms (SpaCy) as well as precomputed models allow implementers to enter a search project "one level up", benefitting from a technology stack which a few years ago would have been unthinkable.

References

[1] Bengio Y et al: Neural Probabilistic Language Models. In: Holmes D.E., Jain L.C. (eds.) Innovations in Machine Learning. Studies in Fuzziness and Soft Computing, 2006, p. 194. Springer, Berlin, Heidelberg DOI: 10.1007/3-540-33486-6_6

[2] https://radimrehurek.com/gensim, retrieved 30.9.2018

[3] https://github.com/Hironsan/awesome-embedding-models, retrieved 30.9.2018

[4] http://ruder.io/word-embeddings-1, retrieved 30.9.2018

[5] Shperber G: A gentle introduction to Doc2Vec, https://medium.com/scaleabout/a-gentle-introduction-to-doc2vec-db3e8c0cce5e, retrieved 30.9.2018

[6] https://luminoso.com, retrieved 30.9.2018

[7] https://nlp.stanford.edu/projects/snli, retrieved 30.9.2018

[8] Kim S, Hong, J-H, Kang I, Kwak N: Semantic sentence matching with densely-connected recurrent and co-attentive Information https://arxiv.org/pdf/1805.11360.pdf

[9] Cavar D, Herring J, Meyer A: Case Law Analysis using Deep NLP and Knowledge Graphs, Proceedings of the LREC 2018, PDF http://lrec-conf.org/workshops/lrec2018/W22/pdf/7_W22.pdf, retrieved 17.10.2018

Stefan Geißler
Expert System
Deutschland GmbH
Blumenstr 15
69115 Heidelberg
Germany
Mobile: +49 174 6595713
skf.geissler@googlemail.com
http://www.linkedin.com/in/stefangeissler

Citation: Geißler S: Trends in scientific document search. Infozine **2018**, Special Issue 2, 38–39
DOI 10.3929/ethz-b-000297324

Published: November 15, 2018